

A stochastic coordinate descent inertial primal-dual algorithm for large-scale composite optimization

Meng Wen^{1,2}, Yu-Chao Tang³, Jigen Peng^{1,2}

1. School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, P.R. China
2. Beijing Center for Mathematics and Information Interdisciplinary Sciences, Beijing, P.R. China
3. Department of Mathematics, NanChang University, Nanchang 330031, P.R. China

Abstract In this paper we consider an inertial primal-dual algorithm to compute the minimizations of the sum of two convex functions and the composition of another convex function with a continuous linear operator. With the idea of coordinate descent, we design a stochastic coordinate descent inertial primal-dual splitting algorithm. Moreover, in order to prove the convergence of the proposed inertial algorithm, we formulate first the inertial version of the randomized Krasnosel'skii-Mann iterations algorithm for approximating the set of fixed points of a nonexpansive operator and investigate its convergence properties. Then the convergence of stochastic coordinate descent inertial primal-dual splitting algorithm is derived by applying the inertial version of the randomized Krasnosel'skii-Mann iterations to the composition of the proximity operator. Finally, we give two applications of our method. (1) In the case of stochastic minibatch optimization, the algorithm can be applicated to split a composite objective function into blocks, each of these blocks being processed sequentially by the computer. (2) In the case of distributed optimization, we consider a set of N networked agents endowed with private cost functions and seeking to find a consensus on the minimizer of the aggregate cost. In that case, we obtain a distributed iterative algorithm where isolated components of the network are activated in an uncoordinated fashion and passing in an asynchronous manner. Numerical results demonstrate the efficiency of the method in the framework of large scale machine learning applications. Generally speaking, our

* Corresponding author.

E-mail address: wen5495688@163.com

method converges faster than existing methods, while keeping the computational cost of each iteration basically unchanged.

Keywords: distributed optimization; large-scale learning; proximity operator; inertial
MR(2000) Subject Classification 47H09, 90C25,

1 Introduction

The purpose of this paper is to designing and discussing an efficient algorithmic framework with inertial version for minimizing the following problem

$$\min_{x \in \mathcal{X}} f(x) + g(x) + (h \circ D)(x), \quad (1.1)$$

where \mathcal{X} and \mathcal{Y} are two finite-dimensional Euclidean spaces, and $n = \dim \mathcal{X}$, $m = \dim \mathcal{Y}$, $f, g \in \Gamma_0(\mathcal{X})$, $h \in \Gamma_0(\mathcal{Y})$, f is differentiable on \mathcal{Y} and $D : \mathcal{X} \rightarrow \mathcal{Y}$ a linear transform. Here and in what follows, for a real Hilbert space \tilde{H} , $\Gamma_0(\tilde{H})$ denotes the collection of all proper lower semi-continuous convex functions from \tilde{H} to $(-\infty, +\infty]$. Despite its simplicity, when $g = 0$ many problems in image processing can be formulated in the form of (1.1).

In this paper, the contributions of us are the following aspects:

(I) We provide a modification of the primal-dual algorithm to solve the general Problem (1.1), which is inspired by the inertial forward-backward splitting method[22]. We refer to our algorithm as IADMM⁺. When $\alpha_k = 0$, the ADMM⁺ algorithm introduced by Bianchi [2] is a special case of our algorithm. In particular, we propose simple and easy to compute diagonal preconditioners for which convergence of the algorithm is guaranteed without the need to compute any step size parameters. we call this algorithm as PADMM⁺.

(II) Based on the results of Bianchi [2] and Radu Ioan et al [9], we introduce the idea of inertial version on randomized krasnoselskii mann iterations. The form of Krasnosel'skii-Mann iterations can be translated into fixed point iterations of a given operator having a contraction-like property. Interestingly, IADMM⁺ is a special instances of the Inertial Krasnosel'skii-Mann iterations. By the view of stochastic coordinate descent, we know that at each iteration, the algorithm is only to update a

random subset of coordinates. Although this leads to a perturbed version of the initial Inertial Krasnosel'skii-Mann iterations, but it can be proved to preserve the convergence properties of the initial unperturbed version. Moreover, stochastic coordinate descent has been used in the literature [11,23-24] for proximal gradient algorithms. We believe that its application to the broader class of Inertial Krasnosel'skii-Mann algorithms can potentially lead to various algorithms well suited to large-scale optimization problems.

(III) We use our views to large-scale optimization problems which arises in signal processing and machine learning contexts. We prove that the general idea of stochastic coordinate descent gives a unified framework allowing to derive stochastic inertial algorithms of different kinds. Furthermore, we give two application examples. Firstly, we propose a new preconditioned stochastic approximation algorithm by applying stochastic coordinate descent on the top of PADMM⁺. The algorithm is called as preconditioned stochastic minibatch primal-dual splitting algorithm (PSMPDS). Secondly, we introduce a random asynchronous distributed optimization methods with preconditioning that we call as preconditioned distributed asynchronous primal-dual splitting algorithm (PDAPDS). The algorithm can be used to efficiently solve an optimization problem over a network of communicating agents. The algorithms are asynchronous in the sense that some components of the network are allowed to wake up at random and perform local updates, while the rest of the network stands still. No coordinator or global clock is needed. The frequency of activation of the various network components is likely to vary.

The rest of this paper is organized as follows. In the next section, we introduce some notations used throughout in the paper. In section 3, we devote to introduce IPDS and IADMM⁺ algorithm, and the relation between them, we also show how the IADMM⁺ includes ADMM⁺ and the Forward-Backward algorithm as special cases. In section 4, we present the preconditioned primal-dual algorithm and give conditions under which convergence of the algorithm is guaranteed. In section 5, we provide our main result on the convergence of Inertial Krasnosel'skii-Mann algorithms with randomized coordinate descent. In section 6, we propose a stochastic approximation algorithm from the PADMM⁺. In section 7, we address the problem of asynchronous distributed optimization. In the final section, we show the numerical performance and efficiency of propose algorithm through some examples in the context of large-scale

l_1 -regularized logistic regression.

2 Preliminaries

Throughout the paper, we denote by $\langle \cdot, \cdot \rangle$ the inner product on \mathcal{X} and by $\|\cdot\|$ the norm on \mathcal{X} .

Assumption 2.1. *The infimum of Problem (1.1) is attained. Moreover, the following qualification condition holds*

$$0 \in \text{ri}(\text{dom } h - D \text{ dom } g).$$

The dual problem corresponding to the primal Problem (1.1) is written

$$\min_{y \in \mathcal{Y}} (f + g)^*(-D^*y) + h^*(y),$$

where a^* denotes the Legendre-Fenchel transform of a function a and where D^* is the adjoint of D . With the Assumption 2.1, the classical Fenchel-Rockafellar duality theory [3], [10] shows that

$$\min_{x \in \mathcal{X}} f(x) + g(x) + (h \circ D)(x) = \min_{y \in \mathcal{Y}} (f + g)^*(-D^*y) + h^*(y). \quad (2.1)$$

Definition 2.1. Let f be a real-valued convex function on \mathcal{X} , the operator prox_f is defined by

$$\begin{aligned} \text{prox}_f : \mathcal{X} &\rightarrow \mathcal{X} \\ x &\mapsto \arg \min_{y \in \mathcal{X}} f(y) + \frac{1}{2} \|x - y\|_2^2, \end{aligned}$$

called the proximity operator of f .

Definition 2.2. Let A be a closed convex set of \mathcal{X} . Then the indicator function of A is defined as

$$\iota_A(x) = \begin{cases} 0, & \text{if } x \in A, \\ \infty, & \text{otherwise.} \end{cases}$$

It can easy see the proximity operator of the indicator function in a closed convex subset A can be reduced a projection operator onto this closed convex set A . That is,

$$prox_{\iota_A} = proj_A,$$

where $proj$ is the projection operator of A .

Definition 2.3. (Nonexpansive operators and firmly nonexpansive operators [3]). Let \mathcal{H} be a Euclidean space (we refer to [3] for an extension to Hilbert spaces). An operator $T : \mathcal{H} \rightarrow \mathcal{H}$ is nonexpansive if and only if it satisfies

$$\|Tx - Ty\|_2 \leq \|x - y\|_2 \text{ for all } (x, y) \in \mathcal{H}^2.$$

T is firmly nonexpansive if and only if it satisfies one of the following equivalent conditions:

- (i) $\|Tx - Ty\|_2^2 \leq \langle Tx - Ty, x - y \rangle$ for all $(x, y) \in \mathcal{H}^2$;
- (ii) $\|Tx - Ty\|_2^2 = \|x - y\|_2^2 - \|(I - T)x - (I - T)y\|_2^2$ for all $(x, y) \in \mathcal{H}^2$.

It is easy to show from the above definitions that a firmly nonexpansive operator T is nonexpansive.

Definition 2.4. A mapping $T : \mathcal{H} \rightarrow \mathcal{H}$ is said to be an averaged mapping, if it can be written as the average of the identity I and a nonexpansive mapping; that is,

$$T = (1 - \alpha)I + \alpha S, \tag{2.2}$$

where α is a number in $]0, 1[$ and $S : \mathcal{H} \rightarrow \mathcal{H}$ is nonexpansive. More precisely, when (2.2) or the following inequality (2.3) holds, we say that T is α -averaged.

$$\|Tx - Ty\|^2 \leq \|x - y\|^2 - \frac{(1 - \alpha)}{\alpha} \|(I - T)x - (I - T)y\|^2, \forall x, y \in \mathcal{H}. \tag{2.3}$$

A 1-averaged operator is said non-expansive. A $\frac{1}{2}$ -averaged operator is said firmly non-expansive.

Definition 2.5. A operator B is said to be single-valued and cocoercive with respect to a linear, selfadjoint and positive definite map L ; that is, for all $x, y \in \mathcal{H}$

$$\langle B(x) - B(y), x - y \rangle \geq \|B(x) - B(y)\|_{L^{-1}}^2 \tag{2.4}$$

where, as usual, we denote $\|x\|_{L^{-1}}^2 = \langle L^{-1}x, x \rangle$. Note that in the most simple case where $L = l \text{ Id}$, $l > 0$, the operator B is $1/l$ co-coercive and hence l -Lipschitz. However, we will later see that in some cases, it makes sense to consider more general L .

We refer the readers to [3] for more details. Let $M : \mathcal{H} \rightarrow \mathcal{H}$ be a set-valued operator. We denote by $\text{ran}(M) := \{v \in \mathcal{H} : \exists u \in \mathcal{H}, v \in Mu\}$ the range of M , by $\text{gra}(M) := \{(u, v) \in \mathcal{H}^2 : v \in Mu\}$ its graph, and by M^{-1} its inverse; that is, the set-valued operator with graph $(v, u) \in \mathcal{H}^2 : v \in Mu$. We define $\text{zer}(M) := \{u \in \mathcal{H} : 0 \in Mu\}$. M is said to be monotone if $\forall (u, u') \in \mathcal{H}^2, \forall (v, v') \in Mu \times Mu', \langle u - u', v - v' \rangle \geq 0$ and maximally monotone if there exists no monotone operator M' such that $\text{gra}(M) \subset \text{gra}(M') \neq \text{gra}(M)$.

The resolvent $(I + M)^{-1}$ of a maximally monotone operator $M : \mathcal{H} \rightarrow \mathcal{H}$ is defined and single-valued on \mathcal{H} and firmly nonexpansive. The subdifferential ∂J of $J \in \Gamma_0(\mathcal{H})$ is maximally monotone and $(I + \partial J)^{-1} = \text{prox}_J$.

Further, let us mention some classes of operators that are used in the paper. The operator A is said to be uniformly monotone if there exists an increasing function $\phi_A : [0; +1) \rightarrow [0; +1]$ that vanishes only at 0, and

$$\langle x - y, u - v \rangle \geq \phi_A(\|x - y\|), \forall (x, u), (y, v) \in \text{gra}(A). \quad (2.5)$$

Prominent representatives of the class of uniformly monotone operators are the strongly monotone operators. Let $\gamma > 0$ be arbitrary. We say that A is γ -strongly monotone, if $\langle x - y, u - v \rangle \geq \gamma\|x - y\|^2$, for all $(x, u), (y, v) \in \text{gra}(A)$.

Lemma 2.1. (*Baillon-Haddad Theorem [3, Corollary 18.16]*). *Let $J : \mathcal{H} \rightarrow \mathcal{R}$ be convex, differentiable on \mathcal{H} and such that $\pi \nabla J$ is nonexpansive, for some $\pi \in]0, +\infty[$. Then ∇J is π -cocoercive; that is, $\pi \nabla J$ is firmly nonexpansive.*

Lemma 2.2. (*Composition of averaged operators [4, Theorem 3]*). *Let $\alpha_1 \in]0, 1[$, $\alpha_2 \in]0, 1[$, $T_1 \in \mathcal{A}(\mathcal{H}, \alpha_1)$, and $T_2 \in \mathcal{A}(\mathcal{H}, \alpha_2)$. Then $T_1 \circ T_2 \in \mathcal{A}(\mathcal{H}, \alpha')$, where*

$$\alpha' := \frac{\alpha_1 + \alpha_2 - 2\alpha_1\alpha_2}{1 - \alpha_1\alpha_2}.$$

Lemma 2.3. (*[13]*). *Let \tilde{H} be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\|\cdot\|$, then*

$$\forall x, y \in \tilde{H}, \forall \alpha \in [0, 1], \|\alpha x + (1 - \alpha)y\|^2 = \alpha\|x\|^2 + (1 - \alpha)\|y\|^2 - \alpha(1 - \alpha)\|x - y\|^2.$$

Lemma 2.4. (see[14-16]). Let $(\varphi^k)_{k \in \mathbb{N}}$; $(\delta_k)_{k \in \mathbb{N}}$ and $(\alpha_k)_{k \in \mathbb{N}}$ be sequences in $[0; +1)$ such that $\varphi^{k+1} \leq \varphi^k + \alpha_k(\varphi^k - \varphi^{k-1}) + \delta_k$ for all $k \geq 1$, $\sum_{k \in \mathbb{N}} \delta_k < +\infty$ and there exists a real number α with $0 \leq \alpha_k \leq \alpha < 1$ for all $k \in \mathbb{N}$. Then the following hold:

- (i) $\sum_{k \geq 1} [\varphi^k - \varphi^{k-1}]_+ < +\infty$, where $[t]_+ = \max\{t, 0\}$;
- (ii) there exists $\varphi^* \in [0; +\infty)$ such that $\lim_{k \rightarrow +\infty} \varphi^k = \varphi^*$.

3 An inertial primal-dual splitting algorithm

3.1 Derivation of the algorithm

In the paper [5], Nesterov proposed a modification of the heavy ball method in order to improve the convergence rate on smooth convex functions. The idea of Nesterov was to use the extrapolated point y^k for evaluating the gradient. Moreover, in order to prove optimal convergence rates of the scheme, the extrapolation parameter α_k must satisfy some special conditions. The scheme is given by:

$$\begin{cases} l^k = x^k + \alpha_k(x^k - x^{k-1}), \\ x^{k+1} = l^k - \bar{\lambda}_k \nabla f(l^k), \end{cases} \quad (3.1)$$

where $\bar{\lambda}_k = 1/L$, there are several choices to define an optimal sequence α_k [5-8].

Recently, for Problem (1.1), Condat [1] considered a primal-dual splitting method as follows:

$$\begin{cases} \tilde{y}^{k+1} = \text{prox}_{\sigma h^*}(y^k + \sigma D x^k), \\ \tilde{x}^{k+1} = \text{prox}_{\tau g}(x^k - \tau \nabla f(x^k) - \tau D^*(2\tilde{y}^{k+1} - y^k)), \\ (x^{k+1}, y^{k+1}) = \rho_k(\tilde{x}^{k+1}, \tilde{y}^{k+1}) + (1 - \rho_k)(x^k, y^k), \end{cases} \quad (3.2)$$

where $\sigma > 0$, $\tau > 0$, $\frac{1}{\tau} - \sigma \|D\|^2 > 0$, $\forall k \in \mathbb{N}$, sequences $\rho_k \in]0, \delta[$, and $\delta = 2 - \frac{1}{2}(\frac{1}{\tau} - \sigma \|D\|^2)^{-1} \in [1, 2[$.

The fixed point characterization provided by Condat [1] suggests solving Problem (1.1) via the fixed point iteration scheme (3.2) for a suitable value of the parameter $\sigma > 0$, $\tau > 0$. This iteration, which is referred to as a primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. A

very natural idea is to combine with the primal-dual splitting method and the heavy ball method, so we obtain the following Algorithm.

Algorithm 1 An inertial primal-dual splitting algorithm(IPDS).

Initialization: Choose $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$, relaxation parameters $(\rho_k)_{k \in \mathbb{N}}$, extrapolation parameter α_k and proximal parameters $\sigma > 0, \tau > 0$.

Iterations ($k \geq 0$): Update x^k, y^k as follows

$$\begin{cases} \xi^k = x^k + \alpha_k(x^k - x^{k-1}), \\ \eta^k = y^k + \alpha_k(y^k - y^{k-1}), \\ \tilde{y}^{k+1} = \text{prox}_{\sigma h^*}(\eta^k + \sigma D\xi^k), \\ \tilde{x}^{k+1} = \text{prox}_{\tau g}(\xi^k - \tau \nabla f(\xi^k) - \tau D^*(2\tilde{y}^{k+1} - \eta^k)), \\ (x^{k+1}, y^{k+1}) = \rho_k(\tilde{x}^{k+1}, \tilde{y}^{k+1}) + (1 - \rho_k)(x^k, y^k). \end{cases}$$

End for

Assume that ∇f is cocoercive with respect to L^{-1} (cf.(2.4)). Then for Algorithm 1, we given the following Theorem.

Theorem 3.1. *Let $\sigma > 0, \tau > 0$, $(\alpha_k)_{k \in \mathbb{N}}$ and the sequences $(\rho_k)_{k \in \mathbb{N}}$, be the parameters of Algorithms 1. Let L be a linear, bounded, selfadjoint and positive definite map defined by (2.4) and that the following hold:*

- (i) $\frac{1}{\tau} - \sigma \|D\|^2 > \frac{\|L\|}{2}$,
 - (ii) $(\alpha_k)_{k \in \mathbb{N}}$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $k \geq 1$ and $\rho, \theta, \hat{\delta} > 0$ are such that $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$ and $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1$.
- Let the sequences (x^k, y^k) be generated by Algorithms 1. Then the sequence $\{x_k\}$ converges to a solution of Problem (1.1).*

We consider the case where D is injective(in particular, it is implicit that $\dim(\mathcal{X}) \leq \dim(\mathcal{Y})$). In the latter case, we denote by $\mathcal{R} = \text{Im}(D)$ the image of D and by D^{-1} the inverse of D on $\mathcal{R} \rightarrow \mathcal{X}$. We emphasize the fact that the inclusion $\mathcal{R} \subset \mathcal{Y}$ might be strict. We make the following assumption:

Assumption 3.1. *The following facts holds true:*

- (1) D is injective;
- (2) $\nabla(f \circ D)^{-1}$ is cocoercive with respect to \bar{L}^{-1} (cf.(2.4)) on \mathcal{R} .

For proximal parameters $\mu > 0$, $\tau > 0$, we consider the following algorithm which we shall refer to as Inertial ADMM⁺ (IADMM⁺).

Algorithm 2 Inertial ADMM⁺ (IADMM⁺).

Iterations ($k \geq 0$): Update x^k , y^k as follows

$$\begin{cases} \xi^k = x^k + \alpha_k(x^k - x^{k-1}), \\ \eta^k = y^k + \alpha_k(y^k - y^{k-1}), \\ z^{k+1} = \arg \min_{w \in \mathcal{Y}} [h(w) + \frac{\|w - (D\xi^k + \mu\eta^k)\|^2}{2\mu}], \end{cases} \quad (3.1a)$$

$$y^{k+1} = \eta^k + \mu^{-1}(D\xi^k - z^{k+1}), \quad (3.2b)$$

$$u^{k+1} = (1 - \tau\mu^{-1})D\xi^k + \tau\mu^{-1}z^{k+1}, \quad (3.3c)$$

$$x^{k+1} = \arg \min_{w \in \mathcal{X}} [g(w) + \langle \nabla f(\xi^k), w \rangle + \frac{\|Dw - u^{k+1} - \tau y^{k+1}\|^2}{2\tau}]. \quad (3.4d)$$

End for

Theorem 3.2. Assume that the minimization Problem (1.1) is consistent, $\mu > 0$, and $\tau > 0$. Let Assumption 2.1 and Assumption 3.1 hold true and \bar{L} be a linear, bounded, selfadjoint and positive definite map defined by (2.4) and $\frac{1}{\tau} - \frac{1}{\mu} > \frac{\|\bar{L}\|}{2}$. Suppose that $(\alpha_k)_{k \in \mathbb{N}}$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $k \geq 1$ and $\rho, \theta, \hat{\delta} > 0$ are such that $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$ and $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1$. Let the sequences (x^k, y^k) be generated by Algorithms 2. Then the sequence $\{x^k\}$ converges to a solution of Problem (1.1).

3.2 Proofs of convergence

From the proof of Theorem 3.1 [1], we know that (3.2) has the structure of a forward-backward iteration, when expressed in terms of nonexpansive operators on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, equipped with a particular inner product.

Let the inner product $\langle \cdot, \cdot \rangle_I$ in \mathcal{Z} be defined as

$$\langle z, z' \rangle := \langle x, x' \rangle + \langle y, y' \rangle, \quad \forall z = (x, y), \quad z' = (x', y') \in \mathcal{Z}.$$

By endowing \mathcal{Z} with this inner product, we obtain the Euclidean space denoted by \mathcal{Z}_I . Let us define the bounded linear operator on \mathcal{Z} ,

$$P : \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} \frac{1}{\tau} & D^* \\ D & \frac{1}{\sigma} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}. \quad (3.4)$$

From the condition (i), we can easily check that P is positive definite. Hence, we can define another inner product $\langle \cdot, \cdot \rangle_P$ and norm $\| \cdot \|_P = \langle \cdot, \cdot \rangle_P^{\frac{1}{2}}$ in \mathcal{Z} as

$$\langle z, z' \rangle_P = \langle z, Pz' \rangle_I. \quad (3.5)$$

We denote by \mathcal{Z}_P the corresponding Euclidean space.

Lemma 3.1. (*[1]*). *Let the conditions (i)-(iii) in Theorem 3.1[1] be true. For every $n \in \mathbb{N}$, the following inclusion is satisfied by $\tilde{z}^{k+1} := (\tilde{x}^{k+1}, \tilde{y}^{k+1})$ computed by (3.2):*

$$\tilde{z}^{k+1} := (I + P^{-1} \circ A)^{-1} \circ (I - P^{-1} \circ B)(z^k), \quad (3.6)$$

where

$$A := \begin{pmatrix} \partial g & D^* \\ -D & \partial h^* \end{pmatrix}, B := \begin{pmatrix} \nabla f \\ 0 \end{pmatrix}.$$

Set $M_1 = P^{-1} \circ A$, $M_2 = P^{-1} \circ B$, $T_1 = (I + M_1)^{-1}$, $T_2 = (I - M_2)^{-1}$, and $T = T_1 \circ T_2$. Then $T_1 \in \mathcal{A}(\mathcal{Z}_P, \frac{1}{2})$ and $T_2 \in \mathcal{A}(\mathcal{Z}_P, \frac{1}{2\kappa})$, $\kappa := (\frac{1}{\tau} - \sigma \|D\|^2)/\beta$. Then $T \in \mathcal{A}(\mathcal{Z}_P, \frac{1}{\delta})$ and $\delta = 2 - \frac{1}{2\kappa}$.

Lemma 3.2. (*[9]*). *Let \tilde{M} be a nonempty closed and affine subset of a Hilbert space $\tilde{\mathcal{H}}$ and $T : \tilde{M} \rightarrow \tilde{M}$ a nonexpansive operator such that $\text{Fix}(T) \neq \emptyset$. Considering the following iterative scheme:*

$$x^{k+1} = x^k + \alpha_k(x^k - x^{k-1}) + \rho_k[T(x^k + \alpha_k(x^k - x^{k-1})) - x^k - \alpha_k(x^k - x^{k-1})], \quad (3.7)$$

where $x^0; x^1$ are arbitrarily chosen in \tilde{M} , $(\alpha_k)_{k \in \mathbb{N}}$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $n \geq 1$ and $\rho, \theta, \hat{\delta} > 0$ are such that $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$ and $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1$.

Then the following statements are true:

- (i) $\sum_{k \in \mathbb{N}} \|x^{k+1} - x^k\|^2 < +\infty$;
- (ii) $(x^k)_{k \in \mathbb{N}}$ converges weakly to a point in $\text{Fix}(T)$.

In association with Lemma 3.1 and Lemma 3.2, we are ready to prove Theorem 3.1

Proof. Set $\nu^k := (\xi^k, \eta^k)$, from (3.6) we can know that the Algorithm 1 can be described as follows:

$$\begin{cases} \nu^k = z^k + \alpha_k(z^k - z^{k-1}), \\ \tilde{z}^{k+1} := (I + P^{-1} \circ A)^{-1} \circ (I - P^{-1} \circ B)(\nu^k). \end{cases} \quad (3.8)$$

Considering the relaxation step, we obtain

$$\begin{cases} \nu^k = z^k + \alpha_k(z^k - z^{k-1}), \\ \tilde{z}^{k+1} := (I + P^{-1} \circ A)^{-1} \circ (I - P^{-1} \circ B)(\nu^k), \\ z^{k+1} := \rho_k(I + P^{-1} \circ A)^{-1} \circ (I - P^{-1} \circ B)(\nu^k) + (1 - \rho_k)\nu^k. \end{cases} \quad (3.9)$$

By Lemma 3.1 we know that $T = T_1 \circ T_2$ is $\frac{1}{\delta}$ -averaged. In particular, it is non-expansive, so from conditions (i)-(ii) and Lemma 3.2 we have that the iterative scheme defined by (3.9) satisfies the following statements:

- (i) $\sum_{k \in \mathbb{N}} \|z^{k+1} - z^k\|^2 < +\infty$;
- (ii) $(z^k)_{k \in \mathbb{N}}$ converges to a point in $Fix(T)$.

Then the sequence $\{x^k\}$ converges to a solution of Problem (1.1). □

Proof of Theorem 3.2 for Algorithm 2. Before providing the proof of Theorem 3.2, let us introduce the following notation and Lemma.

Lemma 3.3. *Given a Euclidean space \mathcal{E} , consider the minimization problem $\min_{\lambda \in \mathcal{E}} \bar{f}(\lambda) + \bar{g}(\lambda) + h(\lambda)$, where $\bar{g}, h \in \Gamma_0(\mathcal{E})$ and where \bar{f} is convex and differentiable on \mathcal{E} and $\nabla \bar{f}$ is cocoercive with respect to L^{-1} (cf. (2.4)). Assume that the infimum is attained and that $0 \in ri(dom h - dom \bar{g})$. Let $\mu > 0$, $\tau > 0$ be such that $\frac{1}{\tau} - \frac{1}{\mu} > \frac{\|\bar{L}\|}{2}$, $(\alpha_k)_{k \in \mathbb{N}}$ be nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $n \geq 1$ and $\rho, \theta, \hat{\delta} > 0$ are such that $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$ and $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1$. Consider the iterates*

$$\begin{cases} \bar{\xi}^k = \lambda^k + \alpha_k(\lambda^k - \lambda^{k-1}), \\ \eta^k = y^k + \alpha_k(y^k - y^{k-1}), \\ y^{k+1} = \text{prox}_{\mu^{-1}h^*}(\eta^k + \mu^{-1}\bar{\xi}^k), \\ \lambda^{k+1} = \text{prox}_{\tau\bar{g}}(\bar{\xi}^k - \tau\nabla\bar{f}(\bar{\xi}^k) - \tau(2y^{k+1} - \eta^k)). \end{cases} \quad (3.10a)$$

Then for any initial value $(\lambda^0, y^0), (\lambda^1, y^1) \in \mathcal{E} \times \mathcal{E}$, the sequence (λ^k, y^k) converges to a primal-dual point $(\tilde{\lambda}, \tilde{y})$, i.e., a solution of the equation

$$\min_{\lambda \in \mathcal{E}} \bar{f}(\lambda) + \bar{g}(\lambda) + h(\lambda) = -\min_{y \in \mathcal{E}} (\bar{f} + \bar{g})^*(y) + h^*(y). \quad (3.11)$$

Proof. It is easy to see that the Lemma 3.3 is a special case of Theorem 3.1. So we can obtain Lemma 3.3 from Theorem 3.1 directly. \square

Elaborating on Lemma 3.3, we are now ready to establish the Theorem 3.2.

By setting $\mathcal{E} = \mathcal{R}$ and by assuming that \mathcal{E} is equipped with the same inner product as \mathcal{Y} , one can notice that the functions $\bar{f} = f \circ D^{-1}$, $\bar{g} = g \circ D^{-1}$ and h satisfy the conditions of Lemma 3.3. Moreover, since $(\bar{f} + \bar{g})^* = (f + g)^* \circ D^*$, one can also notice that (\tilde{x}, \tilde{y}) is a primal-dual point associated with Eq. (2.1) if and only if $(D\tilde{x}, \tilde{y})$ is a primal-dual point associated with Eq. (3.11). With the same idea for the proof of Theorem 1 of [2], we can recover the IADMM⁺ from the iterations (3.10).

3.3 Connections to other algorithms

We will further establish the connections to other existing methods.

When $\alpha_k \equiv 0$, the IADMM⁺ boils down to the ADMM⁺ whose iterations are given by:

$$\begin{cases} z^{k+1} = \operatorname{argmin}_{w \in \mathcal{Y}} [h(w) + \frac{\|w - (Dx^k + \mu y^k)\|^2}{2\mu}], \\ y^{k+1} = y^k + \mu^{-1}(Dx^k - z^{k+1}), \\ u^{k+1} = (1 - \tau\mu^{-1})Dx^k + \tau\mu^{-1}z^{k+1}, \\ x^{k+1} = \operatorname{argmin}_{w \in \mathcal{X}} [g(w) + \langle \nabla f(x^k), w \rangle + \frac{\|Dw - u^{k+1} - \tau y^{k+1}\|^2}{2\tau}]. \end{cases}$$

In the special case $h \equiv 0$, $D = I$ and $\alpha_k \equiv 0$ it can be easily verified that y^k is null for all $k \geq 1$ and $u^k = x^k$. Then, the IADMM⁺ boils down to the standard Forward-Backward algorithm whose iterations are given by:

$$\begin{aligned} x^{k+1} &= \operatorname{argmin}_{w \in \mathcal{X}} g(w) + \frac{1}{2\tau} \|w - (x^k - \tau \nabla f(x^k))\|^2 \\ &= \operatorname{prox}_{\tau g}(x^k - \tau \nabla f(x^k)). \end{aligned}$$

One can remark that μ has disappeared thus it can be set as large as wanted so the condition on stepsize τ from Theorem 3.2 boils down to $\tau < 2/l$. Applications of this algorithm with particular functions appear in well known learning methods such as ISTA [10].

4 Preconditioning

4.1 Convergence of the Preconditioned algorithm

In the context of saddle point problems, Pock and Chambolle [12] proposed a preconditioning of the form

$$P := \begin{pmatrix} \tilde{T}^{-1} & D^* \\ D & \Sigma^{-1} \end{pmatrix}$$

where \tilde{T} and Σ are selfadjoint, positive definite maps. A condition for the positive definiteness of P follows from the following Lemma.

Lemma 4.1. ([12]). *Let A_1, A_2 be symmetric positive definite maps and M a bounded operator. If $\|A_2^{-\frac{1}{2}}MA_1^{-\frac{1}{2}}\| < 1$, then*

$$A := \begin{pmatrix} A_1 & M^* \\ M & A_2 \end{pmatrix}$$

is positive definite.

Now, we study preconditioning techniques for the inertial primal-dual splitting algorithm (IPDS), then we obtain the following algorithm.

Algorithm 3 An inertial primal-dual splitting algorithm with preconditioning (IPDSP).

Initialization: Choose $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$, relaxation parameters $(\rho_k)_{k \in \mathbb{N}}$, extrapolation parameter α_k and positive definite maps \tilde{T}, Σ .

Iterations ($k \geq 0$): Update x^k, y^k as follows

$$\begin{cases} \xi^k = x^k + \alpha_k(x^k - x^{k-1}), \\ \eta^k = y^k + \alpha_k(y^k - y^{k-1}), \\ \tilde{y}^{k+1} = \text{prox}_{\Sigma h^*}(\eta^k + \Sigma D \xi^k), \\ \tilde{x}^{k+1} = \text{prox}_{\tilde{T}g}(\xi^k - \tilde{T} \nabla f(\xi^k) - \tilde{T} D^*(2\tilde{y}^{k+1} - \eta^k)), \\ (x^{k+1}, y^{k+1}) = \rho_k(\tilde{x}^{k+1}, \tilde{y}^{k+1}) + (1 - \rho_k)(x^k, y^k). \end{cases}$$

End for

It turns out that the resulting method converges under appropriate conditions.

Theorem 4.1. *In the setting of Theorem 3.1 let furthermore ∇f be co-coercive w.r.t. a bound, linear, symmetric and positive linear maps E^{-1} . If it holds that*

- (i) $\tilde{T}^{-1} - \frac{1}{2}E > 0$;
 - (ii) $\|\tilde{T}^{-1}\| - \|\Sigma\|\|D\|^2 > \frac{\|E\|}{2}$,
 - (iii) $(\alpha_k)_{k \in \mathbb{N}}$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $n \geq 1$ and $\rho, \theta, \hat{\delta} > 0$ are such that $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$ and $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1$.
- Then the sequence $\{x^k\}$ converges to a solution of Problem (1.1).

Proof. It is easy to check that from the condition (i)-(ii), we can obtain $\|\tilde{T}^{-1} - \frac{1}{2}E\|^{-\frac{1}{2}}\|D\Sigma^{\frac{1}{2}}\| < 1$.

Set

$$C := \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}$$

Then from Lemma 4.1, we can know that $P - \frac{1}{2}C$ is positive definite. Therefore, with the same proof of Theorem 3.1, we can obtain Theorem 4.1. □

For selfadjoint, positive definite maps \tilde{T} , Ψ , we consider the following algorithm which we shall refer to as Preconditioning ADMM⁺(PADMM⁺).

Algorithm 4 Preconditioning ADMM⁺(PADMM⁺).

Iterations ($k \geq 0$): Update x^k, y^k as follows

$$\begin{cases} \xi^k = x^k + \alpha_k(x^k - x^{k-1}), \\ \eta^k = y^k + \alpha_k(y^k - y^{k-1}), \\ z^{k+1} = \arg \min_{w \in \mathcal{Y}} [h(w) + \frac{\|w - (D\xi^k + \Psi\eta^k)\|_{\Psi^{-1}}^2}{2}], \end{cases} \quad (4.1a)$$

$$y^{k+1} = \eta^k + \Psi^{-1}(D\xi^k - z^{k+1}), \quad (4.1b)$$

$$u^{k+1} = (I - \tilde{T}\Psi^{-1})D\xi^k + \tilde{T}\Psi^{-1}z^{k+1}, \quad (4.1c)$$

$$\begin{cases} x^{k+1} = \arg \min_{w \in \mathcal{X}} [g(w) + \langle \nabla f(\xi^k), w \rangle + \frac{\|Dw - u^{k+1} - \tilde{T}y^{k+1}\|_{\tilde{T}^{-1}}^2}{2}]. \end{cases} \quad (4.1d)$$

End for

Theorem 4.2. *In the setting of Theorem 3.2 let furthermore $\nabla \bar{f}$ be co-coercive w.r.t. a bound, linear, symmetric and positive linear maps \bar{E}^{-1} . If it holds that*

$$(i) \tilde{T}^{-1} - \frac{1}{2}\bar{E} > 0;$$

$$(ii) \|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\bar{E}\|}{2},$$

$$(iii) (\alpha_k)_{k \in \mathbb{N}} \text{ is nondecreasing with } \alpha_1 = 0 \text{ and } 0 \leq \alpha_k \leq \alpha < 1 \text{ for every } n \geq 1 \text{ and } \rho, \theta, \hat{\delta} > 0 \text{ are such that } \hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2} \text{ and } 0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \forall k \geq 1.$$

Then the sequence $\{x^k\}$ converges to a solution of Problem (1.1).

Proof. It is easy to check that from the condition (i)-(ii), we can obtain $\|(\tilde{T}^{-1} - \frac{1}{2}\bar{E})^{-\frac{1}{2}}\Psi^{-\frac{1}{2}}\| < 1$.

Set

$$\bar{P} := \begin{pmatrix} \tilde{T}^{-1} & I \\ I & \Psi \end{pmatrix}, \bar{C} := \begin{pmatrix} \bar{E} & 0 \\ 0 & 0 \end{pmatrix}.$$

Then from Lemma 4.1, we can know that $\bar{P} - \frac{1}{2}\bar{C}$ is positive definite. Therefore, with the same proof of Theorem 3.2, we can obtain Theorem 4.2.

□

4.2 Diagonal Preconditioning

In this section, we show how we can choose pointwise step sizes for both the primal and the dual variables that will ensure the convergence of the algorithm. The next result is an adaption of the preconditioner proposed in [11].

Lemma 4.2. *Assume that ∇f is co-coercive with respect to diagonal matrices E^{-1} , where $E = \text{diag}(e_1, \dots, e_n)$. Fix $\gamma \in (0, 2)$, $r > 0$, $s \in [0, 2]$ and let $\tilde{T} = \text{diag}(\tau_1, \dots, \tau_n)$ and $\Psi = \text{diag}(\varphi_1, \dots, \varphi_m)$ with*

$$\tau_j = \frac{1}{\frac{e_j}{\gamma} + \sum_{i=1}^m |D_{i,j}|^{2-s}}, \varphi_i = \frac{1}{r} \sum_{j=1}^n |D_{i,j}|^s, \quad (4.2)$$

then it holds that

$$\tilde{T}^{-1} - \frac{1}{2}E > 0, \Psi > 0, \quad (4.3)$$

$$\|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\bar{E}\|}{2}. \quad (4.4)$$

Proof. The first two conditions follow from the fact that for diagonal matrices, the (4.3) can be written pointwise. By the definition of τ_j , and φ_i it follows that for any $s \in [0, 2]$ and using the convention that $0^0 = 0$,

$$\frac{1}{\tau_j} - \frac{e_j}{2} > \frac{1}{\tau_j} - \frac{e_j}{\gamma} = r \sum_{i=1}^m |D_{i,j}|^{2-s} \geq 0,$$

and

$$\varphi_i = \frac{1}{r} \sum_{j=1}^n |D_{i,j}|^s \geq 0.$$

We will prove (4.4). It is easy to see the proof of (4.4) is equivalent to the proof of

$$\|\Psi^{-\frac{1}{2}} D(\tilde{T}^{-1} - \frac{1}{2}E)^{-\frac{1}{2}}\| < 1. \quad (4.5)$$

So we first show (4.5), For any $s \in [0, 2]$,

$$\begin{aligned} \|\Psi^{-\frac{1}{2}} D(\tilde{T}^{-1} - \frac{1}{2}E)^{-\frac{1}{2}} x\|^2 &= \sum_{i=1}^m \left(\sum_{j=1}^n \frac{1}{\sqrt{\varphi_i}} D_{i,j} \frac{1}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j \right)^2 \\ &= \sum_{i=1}^m \frac{1}{\varphi_i} \left(\sum_{j=1}^n D_{i,j} \frac{1}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j \right)^2 \\ &\leq \sum_{i=1}^m \frac{1}{\varphi_i} \left(\sum_{j=1}^n |D_{i,j}|^{\frac{s}{2}} |D_{i,j}|^{1-\frac{s}{2}} \frac{1}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j \right)^2 \\ &\leq \sum_{i=1}^m \frac{1}{\varphi_i} \left(\sum_{j=1}^n |D_{i,j}|^s \right) \left(\sum_{j=1}^n |D_{i,j}|^{2-s} \frac{1}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j^2 \right). \end{aligned} \quad (4.6)$$

By definition of τ_j and φ_i , and introducing $r > 0$, the above estimate can be simplified to

$$\begin{aligned} &\sum_{i=1}^m \frac{1}{\varphi_i} \left(\sum_{j=1}^n |D_{i,j}|^s \right) \left(\sum_{j=1}^n |D_{i,j}|^{2-s} \frac{r}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j^2 \right) \\ &= \sum_{i=1}^m \sum_{j=1}^n |D_{i,j}|^{2-s} \frac{r}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j^2 \end{aligned}$$

$$\leq \sum_{j=1}^n \left(\sum_{i=1}^m |D_{i,j}|^{2-s} \right) \frac{r}{\sqrt{\frac{1}{\tau_j} - \frac{e_j}{2}}} x_j^2 = \|x\|^2. \quad (4.7)$$

Using the above estimate in the definition of the operator norm, we obtain the desired result

$$\begin{aligned} & \|\Psi^{-\frac{1}{2}} D (\tilde{T}^{-1} - \frac{1}{2} E)^{-\frac{1}{2}}\|^2 \\ &= \sup_{x \neq 0} \frac{\|\Psi^{-\frac{1}{2}} D (\tilde{T}^{-1} - \frac{1}{2} E)^{-\frac{1}{2}} x\|^2}{\|x\|^2} \leq 1. \end{aligned} \quad (4.8)$$

□

Remark 4.1. In particular, for $D = I_Y$, we obtain that

$$\tau_j = \frac{1}{\frac{e_j}{\gamma} + n}, \varphi_i = \frac{1}{r} n, \quad (4.9)$$

then it also holds (4.3) and (4.4).

For $D = I_Y$, from Theorem 4.2, we know that the PADMM⁺ iterates are generated by the action of a nonexpansive operator. Then by Lemma 2.3 we know that a stochastic coordinate descent version of the PADMM⁺ converges towards a primal-dual point. This result will be exploited in two directions: first, we describe a stochastic minibatch algorithm, where a large dataset is randomly split into smaller chunks. Second, we develop an asynchronous version of the PADMM⁺ in the context where it is distributed on a graph.

5 Coordinate descent

5.1 Randomized krasnosel'skii-mann iterations

Consider the space $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_J$ for some $J \in \mathbb{N}^*$ where for any j , \mathcal{Z}_j is a Euclidean space. For \mathcal{Z} equipped with the scalar product $\langle x, y \rangle = \sum_{j=1}^J \langle x_j, y_j \rangle_{\mathcal{Z}_j}$ where $\langle \cdot, \cdot \rangle_{\mathcal{Z}_j}$ is the scalar product in \mathcal{Z}_j . For $j \in \{1, \dots, J\}$, let $T_j : \mathcal{Z} \rightarrow \mathcal{Z}_j$ be the components of the output of operator $T : \mathcal{Z} \rightarrow \mathcal{Z}$ corresponding to \mathcal{Z}_j , so, we have $Tx = (T_1x, \dots, T_Jx)$.

Let $2^{\mathcal{J}}$ be the power set of $\mathcal{J} = \{1, \dots, J\}$. For any $\vartheta \in 2^{\mathcal{J}}$, we denote the operator $\hat{T}^{\vartheta} : \mathcal{Z} \rightarrow \mathcal{Z}$ by $\hat{T}_j^{\vartheta} x = T_j x$ for $j \in \vartheta$ and $\hat{T}_j^{\vartheta} x = x_j$ for otherwise. On some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we introduce a random i.i.d. sequence $(\zeta^k)_{k \in \mathbb{N}^*}$ such that $\zeta^k : \Omega \rightarrow 2^{\mathcal{J}}$ i.e. $\zeta^k(\omega)$ is a subset of \mathcal{J} . Assume that the following holds:

$$\forall j \in \mathcal{J}, \exists \vartheta \in 2^{\mathcal{J}}, j \in \vartheta \quad \text{and} \quad \mathbb{P}(\zeta_1 = \vartheta) > 0. \quad (5.1)$$

Lemma 5.1. (Theorem 3 of [2]). Let $T : \mathcal{Z} \rightarrow \mathcal{Z}$ be \tilde{a} -averaged and $\text{Fix}(T) \neq \emptyset$. Let $(\zeta^k)_{k \in \mathbb{N}^*}$ be a random i.i.d. sequence on $2^{\mathcal{J}}$ such that Condition (5.1) holds. If for all k , sequence $(\rho_k)_{k \in \mathbb{N}}$ satisfies

$$0 < \liminf_{k \rightarrow \infty} \rho_k \leq \limsup_{k \rightarrow \infty} \rho_k < \frac{1}{\tilde{a}}.$$

Then, almost surely, the iterated sequence

$$x^{k+1} = x^k + \rho_k (\hat{T}^{(\zeta^{k+1})} x^k - x^k) \quad (5.2)$$

converges to some point in $\text{Fix}(T)$.

In particular, if T is nonexpansive, and for all k , sequence $(\rho_k)_{k \in \mathbb{N}}$ satisfies

$$0 < \liminf_k \rho_k \leq \limsup_k \rho_k < 1.$$

We can know the iterated sequence (5.2) converges to some point in $\text{Fix}(T)$. Then we obtain the following theorem.

Theorem 5.1. Let $T : \mathcal{Z} \rightarrow \mathcal{Z}$ be nonexpansive and $\text{Fix}(T) \neq \emptyset$. Let $(\zeta^k)_{k \in \mathbb{N}^*}$ be a random i.i.d. sequence on $2^{\mathcal{J}}$ such that Condition (5.1) holds. We consider the following iterative scheme:

$$x^{k+1} = x^k + \alpha_k (x^k - x^{k-1}) + \rho_k [\hat{T}^{(\zeta^{k+1})} (x^k + \alpha_k (x^k - x^{k-1})) - x^k - \alpha_k (x^k - x^{k-1})], \quad (5.3)$$

where $x^0; x^1$ are arbitrarily chosen in \mathcal{Z} , $(\alpha_k)_{k \in \mathbb{N}}$ is nondecreasing with $\alpha_1 = 0$ and $0 \leq \alpha_k \leq \alpha < 1$ for every $k \geq 1$ and $\rho; \theta; \hat{\delta} > 0$ are such that

- (i) $\hat{\delta} > \frac{\alpha^2(1+\alpha)+\alpha\theta}{1-\alpha^2}$;
- (ii) $0 < \rho \leq \rho_k < \frac{\hat{\delta}-\alpha[\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]}{\hat{\delta}[1+\alpha(1+\alpha)+\alpha\hat{\delta}+\theta]} \quad \forall k \geq 1$.

Then, almost surely, the iterated sequence $\{x^k\}$ converges to some point in $\text{Fix}(T)$.

Proof. Let us start with the remark that, due to the choice of $\hat{\delta}$, $\rho_k \in (0, 1)$ for every $k \geq 1$. Set $p_\vartheta = \mathbb{P}(\zeta_1 = \vartheta)$ for any $\vartheta \in 2^{\mathcal{J}}$. Denote by $\|x\|^2 = \langle x, x \rangle$ the squared norm in \mathcal{Z} . Define a new inner product $x \bullet y = \sum_{j=1}^J q_j \langle x_j, y_j \rangle_j$ on \mathcal{Z} where $q_j^{-1} = \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \mathbf{1}_{\{j \in \vartheta\}}$ and let $\|x\|^2 = x \bullet x$ be its associated squared norm. Denote by $w^k = x^k + \alpha_k(x^k - x^{k-1})$. Consider any $\tilde{x} \in \text{Fix}(T)$. It follows from Lemma 2.3 and conditionally to the sigma-field $\mathcal{F}^k = \sigma(\zeta_1, \dots, \zeta^k)$ we have

$$\begin{aligned}
\mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] &= \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|w^k + \rho_k [\hat{T}^{(\zeta^{k+1})} w^k - w^k] - \tilde{x}\|^2 \\
&= (1 - \rho_k) \|w^k - \tilde{x}\|^2 + \rho_k \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - \tilde{x}\|^2 \\
&\quad - \rho_k (1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 \\
&= (1 - \rho_k) \|w^k - \tilde{x}\|^2 + \rho_k \left[\sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \sum_{j \in \vartheta} q_j \|T_j w^k - \tilde{x}_j\|^2 \right. \\
&\quad \left. + \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \sum_{j \notin \vartheta} q_j \|w_j^k - \tilde{x}_j\|^2 \right] \\
&\quad - \rho_k (1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 \\
&= (1 - \rho_k) \|w^k - \tilde{x}\|^2 + \rho_k \|w^k - \tilde{x}\|^2 \\
&\quad + \rho_k \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \sum_{j \in \vartheta} q_j [\|T_j w^k - \tilde{x}_j\|^2 - \|w_j^k - \tilde{x}_j\|^2] \\
&\quad - \rho_k (1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 \\
&= \|w^k - \tilde{x}\|^2 + \rho_k \sum_{j=1}^J [\|T_j w^k - \tilde{x}_j\|^2 - \|w_j^k - \tilde{x}_j\|^2] \\
&\quad - \rho_k (1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 \\
&= \|w^k - \tilde{x}\|^2 + \rho_k [\|T w^k - \tilde{x}\|^2 - \|w^k - \tilde{x}\|^2] \\
&\quad - \rho_k (1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_\vartheta \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2.
\end{aligned}$$

By the the nonexpansiveness of T , we have

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] &\leq \|w^k - \tilde{x}\|^2 \\ &\quad - \rho_k(1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_{\vartheta} \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2.\end{aligned}\tag{5.4}$$

Applying again Lemma 2.3 we have

$$\begin{aligned}\|w^k - \tilde{x}\|^2 &= \|(1 + \alpha_k)(x^k - \tilde{x}) - \alpha_k(x^{k-1} - \tilde{x})\|^2 \\ &= (1 + \alpha_k)\|x^k - \tilde{x}\|^2 - \alpha_k\|x^{k-1} - \tilde{x}\|^2 + \alpha_k(1 + \alpha_k)\|x^k - x^{k-1}\|^2,\end{aligned}$$

hence by (5.4) we obtain

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] &- (1 + \alpha_k)\|x^k - \tilde{x}\|^2 + \alpha_k\|x^{k-1} - \tilde{x}\|^2 \\ &\leq -\rho_k(1 - \rho_k) \sum_{\vartheta \in 2^{\mathcal{J}}} p_{\vartheta} \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 + \alpha_k(1 + \alpha_k)\|x^k - x^{k-1}\|^2.\end{aligned}\tag{5.5}$$

Further, we have

$$\begin{aligned}\sum_{\vartheta \in 2^{\mathcal{J}}} p_{\vartheta} \|\hat{T}^{(\zeta^{k+1})} w^k - w^k\|^2 &= \left\| \frac{1}{\rho_k}(x^{k+1} - x^k) + \frac{\alpha_k}{\rho_k}(x^{k-1} - x^k) \right\|^2 \\ &\geq \frac{1}{\rho_k^2} \|x^{k+1} - x^k\|^2 + \frac{\alpha_k^2}{\rho_k^2} \|x^{k-1} - x^k\|^2 \\ &\quad + \frac{\alpha_k}{\rho_k^2} (-\lambda_k \|x^{k+1} - x^k\|^2 - \frac{1}{\lambda_k} \|x^{k-1} - x^k\|^2),\end{aligned}\tag{5.6}$$

where we denote $\lambda_k = \frac{1}{\alpha_k + \delta \rho_k}$.

We derive from (5.5) and (5.6) the inequality (notice that $\rho_k \in (0; 1)$)

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] &- (1 + \alpha_k)\|x^k - \tilde{x}\|^2 + \alpha_k\|x^{k-1} - \tilde{x}\|^2 \\ &\leq \frac{(1 - \rho_k)(\alpha_k \lambda_k - 1)}{\rho_k} \|x^{k+1} - x^k\|^2 + \gamma_k \|x^k - x^{k-1}\|^2,\end{aligned}\tag{5.7}$$

where

$$\gamma_k := \alpha_k(1 + \alpha_k) + \alpha_k(1 - \rho_k) \frac{1 - \lambda_k \alpha_k}{\lambda_k \rho_k} > 0. \quad (5.8)$$

Taking again into account the choice of λ_k we have

$$\hat{\delta} = \frac{1 - \lambda_k \alpha_k}{\lambda_k \rho_k},$$

and by (5.8) it follows

$$\gamma_k := \alpha_k(1 + \alpha_k) + \alpha_k(1 - \rho_k) \hat{\delta} \leq \alpha(1 + \alpha) + \alpha \hat{\delta}, \forall k \geq 1. \quad (5.9)$$

In the following we use some techniques from [14] adapted to our setting. We define the sequences $\varphi^k := \|x^k - \tilde{x}\|^2$ for all $k \in \mathbb{N}$ and $\varpi^k := \varphi^k - \alpha_k \varphi^{k-1} + \gamma_k \|x^k - x^{k-1}\|^2$, for all $k \geq 1$. Using the monotonicity of $(\alpha_k)_{k \geq 1}$ and the fact that $\varphi^k > 0$ for all $k \in \mathbb{N}$, we get

$$\varpi^{k+1} - \varpi^k \leq \varphi^{k+1} - (1 + \alpha_k) \varphi^k + \alpha_k \varphi^{k-1} + \gamma_{k+1} \|x^{k+1} - x^k\|^2 - \gamma_k \|x^k - x^{k-1}\|^2,$$

which gives by (5.7)

$$\varpi^{k+1} - \varpi^k \leq \left(\frac{(1 - \rho_k)(\alpha_k \lambda_k - 1)}{\rho_k} + \gamma_{k+1} \right) \|x^{k+1} - x^k\|^2, \forall k \geq 1. \quad (5.10)$$

We claim that

$$\frac{(1 - \rho_k)(\alpha_k \lambda_k - 1)}{\rho_k} + \gamma_{k+1} \leq -\theta, \forall k \geq 1. \quad (5.11)$$

Let be $k \geq 1$. Indeed, by the choice of λ_k , we get

$$\begin{aligned} & \frac{(1 - \rho_k)(\alpha_k \lambda_k - 1)}{\rho_k} + \gamma_{k+1} \leq -\theta \\ \iff & \rho_k(\gamma_{k+1} + \theta) + (\alpha_k \lambda_k - 1)(1 - \rho_k) \leq 0, \\ \iff & \rho_k(\gamma_{k+1} + \theta) + \frac{\hat{\delta} \rho_k (1 - \rho_k)}{\alpha_k + \hat{\delta} \rho_k} \leq 0, \\ \iff & (\alpha_k + \hat{\delta} \rho_k)(\gamma_{k+1} + \theta) + \hat{\delta} \rho_k \leq \hat{\delta}. \end{aligned}$$

Thus, by using (5.9), we have

$$(\alpha_k + \hat{\delta}\rho_k)(\gamma_{k+1} + \theta) + \hat{\delta}\rho_k \leq (\alpha_k + \hat{\delta}\rho_k)(\alpha(1 + \alpha) + \alpha\hat{\delta} + \theta) + \hat{\delta}\rho_k \leq \hat{\delta},$$

where the last inequality follows by taking into account the upper bound considered for $(\rho_k)_{k \in \mathbb{N}}$ in (ii). Hence the claim in (5.11) is true.

We obtain from (5.10) and (5.11) that

$$\varpi^{k+1} - \varpi^k \leq -\theta \|x^{k+1} - x^k\|^2, \forall k \geq 1. \quad (5.12)$$

The sequence $(\varpi_k)_{k \geq 1}$ is nonincreasing and the bound for $(\alpha_k)_{k \geq 1}$ delivers

$$-\alpha\varphi^{k-1} \leq \varphi^k - \alpha\varphi^{k-1} \leq \varpi^k \leq \varpi^1, \forall k \geq 1. \quad (5.13)$$

We obtain

$$\varphi^k \leq \alpha^k \varphi^0 + \varpi^1 \sum_{n=0}^{k-1} \alpha^n \leq \alpha^k \varphi^0 + \frac{\varpi^1}{1 - \alpha}, \forall k \geq 1,$$

where we notice that $\varpi^1 = \varphi^1 \geq 0$ (due to the relation $\alpha_1 = 0$). Combining (5.12) and (5.13), we get for all $k \geq 1$

$$\theta \sum_{n=1}^k \|x^{n+1} - x^n\|^2 \leq \varpi^1 - \varpi^{k+1} \leq \varpi^1 + \alpha\varphi^k \leq \alpha^k \varphi^0 + \frac{\varpi^1}{1 - \alpha},$$

which shows that $\sum_{k \in \mathbb{N}} \|x^{k+1} - x^k\|^2 < +\infty$ with respect to the filtration (\mathcal{F}^k) . By $w^k = x^k + \alpha_k(x^k - x^{k-1})$, (5.9) and Lemma 2.4 we derive that $\|x^k - \tilde{x}\|$ converges with probability one towards a random variable that is finite almost everywhere.

Given a countable dense subset Z of $Fix(T)$, there is a probability one set on which $\|x^k - x\| \rightarrow X_x \in [0, \infty)$ for all $x \in Z$. Let $x \in Fix(T)$, let $\varepsilon > 0$, and choose $\tilde{x} \in Z$ such that $\|\tilde{x} - x\| \leq \varepsilon$. With probability one, we have

$$\|x^k - \tilde{x}\| \leq \|x^k - x\| + \|\tilde{x} - x\| \leq X_x + 2\varepsilon,$$

for k large enough. Similarly $\|x^k - \tilde{x}\| \geq X_x - 2\varepsilon$, for k large enough. Therefore, we have

A₁: There is a probability one set on which $\|x^k - \tilde{x}\|$ converges for every $\tilde{x} \in Fix(T)$.

By the definition of w^k and the upper bound requested for $(\alpha_k)_{k \geq 1}$, we get there is a probability one set on which $\|w^k - \tilde{x}\|$ converges for every $\tilde{x} \in \text{Fix}(T)$. On the other hand, with the same proof of Theorem 3 of [2], we know that

$$\begin{aligned} \mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] &\leq \|w^k - \tilde{x}\|^2 \\ &\quad - \rho_k(1 - \rho_k)\|(I - T)w^k\|^2. \end{aligned} \quad (5.14)$$

From the assumption on ρ_k , we know that

$$\begin{aligned} \rho(1 - \frac{\hat{\delta} - \alpha[\alpha(1 + \alpha) + \alpha\hat{\delta} + \theta]}{\hat{\delta}[1 + \alpha(1 + \alpha) + \alpha\hat{\delta} + \theta]})\|(I - T)w^k\|^2 &\leq \rho_k(1 - \rho_k)\|(I - T)w^k\|^2 \\ &\leq \|w^k - \tilde{x}\|^2 - \mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] \\ &\leq \alpha_k^2 \|x^k - x^{k-1}\|^2 \\ &\quad - \mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k] \\ &\leq \|x^k - x^{k-1}\|^2 - \mathbb{E}[\|x^{k+1} - \tilde{x}\|^2 | \mathcal{F}^k]. \end{aligned} \quad (5.15)$$

Taking the expectations on both sides of inequality (5.15) and iterating over k , we obtain

$$\mathbb{E}\|(I - T^k)w^k\|^2 \leq \frac{1}{\rho(1 - \frac{\hat{\delta} - \alpha[\alpha(1 + \alpha) + \alpha\hat{\delta} + \theta]}{\hat{\delta}[1 + \alpha(1 + \alpha) + \alpha\hat{\delta} + \theta]})}(x^0 - \tilde{x})^2.$$

By Markov's inequality and Borel Cantelli's lemma, we therefore obtain:

$$\mathbf{A}_2: (I - T)w^k \rightarrow 0 \text{ almost surely.}$$

We now consider an elementary event in the probability one set where \mathbf{A}_1 and \mathbf{A}_2 hold. On this event, since the sequence $\|w^k - \tilde{x}\|$ converges for $\tilde{x} \in \text{Fix}(T)$, the sequence $(w^k)_{k \in \mathbb{N}}$ is bounded. Since T is nonexpansive, it is continuous, and \mathbf{A}_2 shows that all the accumulation points of $(w^k)_{k \in \mathbb{N}}$ are in $\text{Fix}(T)$. It remains to show that these accumulation points reduce to one point. Assume that x^* is an accumulation point. By \mathbf{A}_1 , $\|w^k - x^*\|$ converges. Therefore, $\lim \|w^k - x^*\| = \liminf \|w^k - x^*\| = 0$, which shows that x^* is unique. □

6 Application to stochastic approximation

6.1 Problem setting

Given an integer $N > 1$, consider the problem of minimizing a sum of composite functions

$$\inf_{x \in \mathcal{X}} \sum_{n=1}^N (f_n(x) + g_n(x)), \quad (6.1)$$

where we make the following assumption:

Assumption 6.1. *For each $n = 1, \dots, N$,*

- (1) *f_n is a convex differentiable function on \mathcal{X} , and its gradient ∇f_n be co-coercive w.r.t. a bound, linear, symmetric and positive linear maps \hat{E}^{-1} ;*
- (2) *$g_n \in \Gamma_0(\mathcal{X})$;*
- (3) *The infimum of Problem (6.1) is attained;*
- (4) *$\cap_{n=1}^N \text{ridom} g_n \neq \emptyset$.*

This problem arises for instance in large-scale learning applications where the learning set is too large to be handled as a single block. Stochastic minibatch approaches consist in splitting the data set into N chunks and to process each chunk in some order, one at a time. The quantity $f_n(x) + g_n(x)$ measures the inadequacy between the model (represented by parameter x) and the n -th chunk of data. Typically, f_n stands for a data fitting term whereas g_n is a regularization term which penalizes the occurrence of erratic solutions. As an example, the case where f_n is quadratic and g_n is the l_1 -norm reduces to the popular LASSO problem [13]. In particular, it is also useful to recover sparse signal.

6.2 Instantiating the PADMM⁺

We regard our stochastic minibatch algorithm as an instance of the PADMM⁺ coupled with a randomized coordinate descent. In order to end that, we rephrase Problem (6.1) as

$$\inf_{x \in \mathcal{X}^N} \sum_{n=1}^N (f_n(x) + g_n(x)) + \iota_{\mathcal{C}}(x), \quad (6.2)$$

where the notation x_n represents the n -th component of any $x \in \mathcal{X}^N$, \mathcal{C} is the space of vectors $x \in \mathcal{X}^N$ such that $x_1 = \dots = x_N$. On the space \mathcal{X}^N , we set $f(x) = \sum_n f_n(x_n)$, $g(x) = \sum_n g_n(x_n)$, $h(x) = \iota_{\mathcal{C}}$ and $D = I_{\mathcal{X}^N}$ the identity matrix. Problem (6.2) is equivalent to

$$\min_{x \in \mathcal{X}^N} f(x) + g(x) + (h \circ D)(x). \quad (6.3)$$

We define the natural scalar product on \mathcal{X}^N as $\langle x, y \rangle = \sum_{n=1}^N \langle x_n, y_n \rangle$. Applying the PADMM⁺ to solve Problem (6.3) leads to the following iterative scheme:

$$\begin{aligned} \eta^k &= y^k + \alpha_k(y^k - y^{k-1}), \\ \xi^k &= x^k + \alpha_k(x^k - x^{k-1}), \\ z^{k+1} &= \text{proj}_{\mathcal{C}}(\xi^k + \Psi\eta^k), \\ y_n^{k+1} &= \eta_n^k + \Psi^{-1}(\xi_n^k - z_n^{k+1}), \\ u_n^{k+1} &= (I - \tilde{T}\Psi^{-1})\xi_n^k + \tilde{T}\Psi^{-1}z_n^{k+1}, \\ x_n^{k+1} &= \arg \min_{w \in \mathcal{X}} [g_n(w) + \langle \nabla f_n(\xi^k), w \rangle + \frac{\|w - u_n^{k+1} - \tilde{T}y_n^{k+1}\|_{\tilde{T}^{-1}}^2}{2}], \end{aligned}$$

where T and Ψ are two diagonal matrices which have same dimensional, $\text{proj}_{\mathcal{C}}$ is the orthogonal projection onto \mathcal{C} . Observe that for any $x \in \mathcal{X}^N$, $\text{proj}_{\mathcal{C}}(x)$ is equivalent to $(\bar{x}, \dots, \bar{x})$ where \bar{x} is the average of vector x , that is $\bar{x} = N^{-1} \sum_n x_n$. Consequently, the components of z^{k+1} are equal and coincide with $\bar{\xi}^k + \Psi\bar{\eta}^k$ where $\bar{\xi}^k$ and $\bar{\eta}^k$ are the averages of ξ^k and η^k respectively. By inspecting the y^k n -update equation above, we notice that the latter equality simplifies even further by noting that $\bar{y}^{k+1} = 0$ or, equivalently, $\bar{\eta}^k = 0$ for all $k \geq 1$ if the algorithm is started with $\bar{y}^0 = 0$. Finally, for any n and $k \geq 1$, the above iterations reduce to

$$\begin{aligned} \eta^k &= y^k + \alpha_k(y^k - y^{k-1}), \\ \xi^k &= x^k + \alpha_k(x^k - x^{k-1}), \\ \bar{\xi}^k &= \frac{1}{N} \sum_{n=1}^N \xi_n^k, \\ y_n^{k+1} &= \eta_n^k + \Psi^{-1}(\xi_n^k - \bar{\xi}^k), \\ u_n^{k+1} &= (I - \tilde{T}\Psi^{-1})\xi_n^k + \tilde{T}\Psi^{-1}\bar{\xi}^k, \end{aligned}$$

$$x_n^{k+1} = \text{prox}_{\tilde{T}g_n}[u_n^{k+1} - \tilde{T}(\nabla f_n(\xi_n^k) + y_n^{k+1})].$$

These iterations can be written more compactly as

Algorithm 5 Minibatch PADMM⁺.

Initialization: Choose $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$, s.t. $\sum_n y_n^0 = 0$.

Do

- $\eta^k = y^k + \alpha_k(y^k - y^{k-1})$,
 $\xi^k = x^k + \alpha_k(x^k - x^{k-1})$,
 $\bar{\xi}^k = \frac{1}{N} \sum_{n=1}^N \xi_n^k$,
 - For batches $n = 1, \dots, N$, do (6.4)
 $y_n^{k+1} = \eta_n^k + \Psi^{-1}(\xi_n^k - \bar{\xi}^k)$,
 $x_n^{k+1} = \text{prox}_{\tilde{T}g_n}[(I - 2\tilde{T}\Psi^{-1})\xi_n^k - \tilde{T}\nabla f_n(\xi_n^k) + 2\tilde{T}\Psi^{-1}\bar{\xi}^k - \tilde{T}\eta_n^k]$.
 - Increment k .
-

The following result is a straightforward consequence of Theorem 4.2.

Theorem 6.1. *Assume that the minimization Problem (6.3) is consistent, T and Ψ are two diagonal matrices which have same dimensional. Let Assumption 6.1 hold true and $\tilde{T}^{-1} - \frac{1}{2}\hat{E} > 0$, $\|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\hat{E}\|}{2}$. Let the sequences (\bar{x}^k, y^k) be generated by Minibatch PADMM⁺. Then for any initial point $(x^0, y^0), (x^1, y^1)$ such that $\bar{y}^0 = 0$, the sequence $\{\bar{x}^k\}$ converges to a solution of Problem (6.3).*

At each step k , the iterations given above involve the whole set of functions $f_n, g_n (n = 1, \dots, N)$. Our aim is now to propose an algorithm which involves a single couple of functions (f_n, g_n) per iteration.

6.3 A stochastic minibatch primal-dual splitting algorithm with preconditioning

We are now in position to state the main algorithm of this section. The proposed preconditioned stochastic minibatch primal-dual splitting algorithm (PSMPDS) is obtained upon applying the randomized coordinate descent on the minibatch PADMM⁺:

Algorithm 6 PSMPDS.

Initialization: Choose $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$.

Do

- Define $\eta^k = y^k + \alpha_k(y^k - y^{k-1})$,
 $\xi^k = x^k + \alpha_k(x^k - x^{k-1})$,
 $\bar{\xi}^k = \frac{1}{N} \sum_{n=1}^N \xi_n^k$, $\bar{\eta}^k = \frac{1}{N} \sum_{n=1}^N \eta_n^k$,
 - Pick up the value of ζ^{k+1} ,
 - For batch $n = \zeta^{k+1}$, set
 $y_n^{k+1} = \eta_n^k - \bar{\eta}^k + \Psi^{-1}(\xi_n^k - \bar{\xi}^k)$, (6.5a)
 $x_n^{k+1} = \text{prox}_{\tilde{T}g_n}[(I - 2\tilde{T}\Psi^{-1})\xi_n^k - \tilde{T}\nabla f_n(\xi_n^k) - \tilde{T}\eta_n^k + 2\tilde{T}(\Psi^{-1}\bar{\xi}^k + \bar{\eta}^k)]$. (6.5b)
 - For all batches $n \neq \zeta^{k+1}$, $y_n^{k+1} = \eta_n^k$, $x_n^{k+1} = \xi_n^k$.
 - Increment k .
-

Assumption 6.2. The random sequence $(\zeta^k)_{k \in \mathbb{N}^*}$ is i.i.d. and satisfies $\mathbb{P}[\zeta^1 = n] > 0$ for all $n = 1, \dots, N$.

Theorem 6.2. Assume that the minimization Problem (6.3) is consistent, T and Ψ are two diagonal matrices which have same dimensional. Let Assumption 6.1 and Assumption 6.2 hold true and $\tilde{T}^{-1} - \frac{1}{2}\hat{E} > 0$, $\|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\hat{E}\|}{2}$. Then for any initial point $(x^0, y^0), (x^1, y^1)$, the sequence $\{\bar{x}^k\}$ generated by PSMPDS algorithm converges to a solution of Problem (6.3).

Proof. Let us define $(\bar{f}, \bar{g}, h, D) = (f, g, h, I_{x^N})$ where the functions f , g , and h are the ones defined in section 6.2. If we replace T, Ψ by μ, τ , then the iterates $((y_n^{k+1})_{n=1}^N, (x_n^{k+1})_{n=1}^N)$ described by Equations (6.4) coincide with the iterates (y^{k+1}, x^{k+1}) described by Equations (3.10). If we write these equations more compactly as $(y^{k+1}, x^{k+1}) = T(\xi^k, \eta^k)$ where $(\xi^k, \eta^k) = (x^k, y^k) + \alpha_k[(x^k, y^k) - (x^{k-1}, y^{k-1})]$, and the operator T acts in the space $\mathcal{Z} = \mathcal{X}^N \times \mathcal{X}^N$, then from the proof of Lemma 3.1, we know that T is \tilde{a} -averaged, where $\tilde{a} = (2 - a_1)^{-1}$ and $a_1 = \frac{\|\hat{E}\|}{2}(\|\tilde{T}^{-1}\| - \|\Psi^{-1}\|)^{-1}$. Defining the selection operator \mathcal{S}_n on \mathcal{Z} as $\mathcal{S}_n(y, x) = (y_n, x_n)$, we obtain that $\mathcal{Z} = \mathcal{S}_1(\mathcal{Z}) \times \dots \times \mathcal{S}_N(\mathcal{Z})$ up to an element reordering. To be compatible with the notations of Section 5.1, we assume that $J = N$ and that the random sequence ζ^k driving the PSMPDS algorithm is set

valued in $\{\{1\}, \dots, \{N\}\} \subset 2^{\mathcal{J}}$. In order to establish Theorem 6.2, we need to show that the iterates (y^{k+1}, x^{k+1}) provided by the PSMPDS algorithm are those who satisfy the equation $(y^{k+1}, x^{k+1}) = T^{(\zeta^{k+1})}(y^k, x^k)$. By the direct application of Theorem 5.1, we can obtain Theorem 6.2.

Let us start with the y -update equation. Since $h = \iota_C$, its Legendre-Fenchel transform is $h^* = \iota_{C^\perp}$ where C^\perp is the orthogonal complement of C in \mathcal{X}^N . Consequently, if we write $(\zeta^{k+1}, v^{k+1}) = T(y^k, x^k)$, and replace μ, τ by T, Ψ then by Eq. (3.10a),

$$\zeta_n^{k+1} = \eta_n^k - \bar{\eta}^k + \Psi^{-1}(\xi_n^k - \bar{\xi}^k) \quad n = 1, \dots, N.$$

Observe that in general, $\bar{y}^k \neq 0$ because in the PSMPDS algorithm, only one component is updated at a time. If $\{n\} = \zeta^{k+1}$, then $y_n^{k+1} = \zeta_n^{k+1}$ which is Eq. (6.5a). All other components of y^k are carried over to y^{k+1} .

By Equation (3.10b) we also get

$$v_n^{k+1} = \text{prox}_{\tilde{T}g_n}[\xi_n^k - \tilde{T}\nabla f_n(\xi_n^k) - \tilde{T}(2y_n^{k+1} - \eta^k)].$$

If $\{n\} = \zeta^{k+1}$, then $x_n^{k+1} = v_n^{k+1}$ can easily be shown to be given by (6.5b). □

7 Distributed optimization

Consider a set of $N > 1$ computing agents that cooperate to solve the minimization Problem (6.1). Here, f_n, g_n are two private functions available at Agent n . Our purpose is to introduce a random distributed algorithm to solve (6.1). The algorithm is asynchronous in the sense that some components of the network are allowed to wake up at random and perform local updates, while the rest of the network stands still. No coordinator or global clock is needed. The frequency of activation of the various network components is likely to vary.

The examples of this problem appear in learning applications where massive training data sets are distributed over a network and processed by distinct machines [17], [18], in resource allocation problems for communication networks [19], or in statistical estimation problems by sensor networks [20], [21].

7.1 Network model and problem formulation

We consider the network as a graph $G = (Q, E)$ where $Q = \{1, \dots, N\}$ is the set of agents/nodes and $E \subset \{1, \dots, N\}^2$ is the set of undirected edges. We write $n \sim m$ whenever $n, m \in E$. Practically, $n \sim m$ means that agents n and m can communicate with each other.

Assumption 7.1. *G is connected and has no self loop.*

Now we introduce some notations. For any $x \in \mathcal{X}^{|Q|}$, we denote by x_n the components of x , i.e., $x = (x_n)_{n \in Q}$. We regard the functions f and g on $\mathcal{X}^{|Q|} \rightarrow (-\infty, +\infty]$ as $f(x) = \sum_{n \in Q} f_n(x_n)$ and $g(x) = \sum_{n \in Q} g_n(x_n)$. So the Problem (6.1) is equal to the minimization of $f(x) + g(x)$ under the constraint that all components of x are equal.

Next we write the latter constraint in a way that involves the graph G . We replace the global consensus constraint by a modified version of the function ι_C . The purpose of us is to ensure global consensus through local consensus over every edge of the graph.

For any $\epsilon \in E$, say $\epsilon = \{n, m\} \in Q$, we define the linear operator $D_\epsilon(x) : \mathcal{X}^{|Q|} \rightarrow \mathcal{X}^2$ as $D_\epsilon(x) = (x_n, x_m)$ where we assume some ordering on the nodes to avoid any ambiguity on the definition of D . We construct the linear operator $D : \mathcal{X}^{|Q|} \rightarrow \mathcal{Y} \triangleq \mathcal{X}^{2|E|}$ as $D(x) = (D_\epsilon(x))_{\epsilon \in E}$ where we also assume some ordering on the edges. Any vector $y \in \mathcal{Y}$ will be written as $y = (y_\epsilon)_{\epsilon \in E}$ where, writing $\epsilon = \{n, m\} \in E$, the component y_ϵ will be represented by the couple $y_\epsilon = (y_\epsilon(n), y_\epsilon(m))$ with $n < m$. We also introduce the subspace of \mathcal{X}^2 defined as $\mathcal{C}_2 = \{(x, x) : x \in \mathcal{X}\}$. Finally, we define $h : \mathcal{Y} \rightarrow (-\infty, +\infty]$ as

$$h(y) = \sum_{\epsilon \in E} \iota_{\mathcal{C}_2}(y_\epsilon). \quad (7.1)$$

Then we consider the following problem:

$$\min_{x \in \mathcal{X}^{|Q|}} f(x) + g(x) + (h \circ D)(x). \quad (7.2)$$

Lemma 7.1. ([2]). *Let Assumptions 7.1 hold true. The minimizers of (7.2) are the tuples (x^*, \dots, x^*) where x^* is any minimizer of (6.1).*

7.2 Instantiating the PADMM⁺

Now we use the PADMM⁺ to solve the Problem (7.2). Since the newly defined function h is separable with respect to the $(y_\epsilon)_{\epsilon \in E}$, we get

$$\text{prox}_{\tilde{T}h}(y) = (\text{prox}_{\tilde{T}h_{C_2}}(y_\epsilon))_{\epsilon \in E} = ((\bar{y}_\epsilon, \bar{y}_\epsilon))_{\epsilon \in E},$$

where $\bar{y}_\epsilon = (y_\epsilon(n) + y_\epsilon(m))/2$ if $\epsilon = \{n, m\}$. With this at hand, the update equation (4.1a) of the PADMM⁺ can be written as

$$z^{k+1} = ((\bar{z}_\epsilon^{k+1}, \bar{z}_\epsilon^{k+1}))_{\epsilon \in E},$$

where

$$\bar{z}^{k+1} = \frac{\xi_n^k + \xi_m^k}{2} + \frac{\Psi(\eta_\epsilon^k(n) + \eta_\epsilon^k(m))}{2},$$

for any $\epsilon = \{n, m\} \in E$. Plugging this equality into Eq. (4.1b) of the PADMM⁺, it can be seen that $\eta_\epsilon^k(n) = -\eta_\epsilon^k(m)$. Therefore

$$\bar{z}^{k+1} = \frac{\xi_n^k + \xi_m^k}{2},$$

for any $k \geq 1$. Moreover

$$y_\epsilon^{k+1} = \frac{\Psi^{-1}(\xi_n^k - \xi_m^k)}{2} + \eta_\epsilon^k(n).$$

Observe that the n -th component of the vector D^*Dx coincides with $d_n x_n$, where d_n is the degree (i.e., the number of neighbors) of node n . From (4.1d) of the PADMM⁺, the n^{th} component of x^{k+1} can be written

$$x_n^{k+1} = \text{prox}_{\tilde{T}g_n/d_n} \left[\frac{(D^*(u^{k+1} - \tilde{T}y^{k+1}))_n - \tilde{T}\nabla f_n(\xi_n^k)}{d_n} \right],$$

where for any $y \in \mathcal{Y}$,

$$(D^*y)_n = \sum_{m: \{n, m\} \in E} y_{\{n, m\}}(n)$$

is the n -th component of $D^*y \in \mathcal{X}^{|Q|}$. Plugging Eq. (4.1c) of the PADMM⁺ together with the expressions of $\bar{z}_{\{n, m\}}^{k+1}$ and $y_{\{n, m\}}^{k+1}$ in the argument of $\text{prox}_{\tilde{T}g_n/d_n}$, we can have

$$x_n^{k+1} = \text{prox}_{\tilde{T}g_n/d_n} \left[(I - \tilde{T}\Psi^{-1})\xi_n^k - \frac{\tilde{T}}{d_n} \nabla f_n(\xi_n^k) + \frac{\tilde{T}}{d_n} \sum_{m: \{n, m\} \in E} (\Psi^{-1}\xi_m^k - \eta_{\{n, m\}}^k(n)) \right].$$

The algorithm is finally described by the following procedure: Prior to the clock tick $k + 1$, the node n has in its memory the variables x_n^k , $\{y_{\{n,m\}}^k(n)\}_{m \sim n}$, and $\{x_m^k\}_{m \sim n}$.

Algorithm 7 Distributed PADMM⁺.

Initialization: Choose $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$, s.t. $\sum_n y_n^0 = 0$.

Do

- Define $\eta^k = y^k + \alpha_k(y^k - y^{k-1})$,
 $\xi^k = x^k + \alpha_k(x^k - x^{k-1})$,
 - For any $n \in Q$, Agent n performs the following operations :

$$y_{\{n,m\}}^{k+1}(n) = \eta_{\{n,m\}}^k(n) + \frac{\xi_n^k - \xi_m^k}{2}, \quad \text{for all } m \sim n, \quad (7.3a)$$

$$x_n^{k+1} = \text{prox}_{\tilde{T}g_n/d_n}[(I - \tilde{T}\Psi^{-1})\xi_n^k - \frac{\tilde{T}}{d_n}\nabla f_n(\xi_n^k) + \frac{\tilde{T}}{d_n}\sum_{m:\{n,m\} \in E}(\Psi^{-1}\xi_m^k - \eta_{\{n,m\}}^k(n))]. \quad (7.3b)$$
 - Agent n sends the parameter y_n^{k+1}, x_n^{k+1} to their neighbors respectively.
 - Increment k .
-

Theorem 7.1. Assume that the minimization Problem (6.1) is consistent, T and Ψ are two diagonal matrices which have same dimensional. Let Assumption 6.1 and Assumption 7.1 hold true and $\tilde{T}^{-1} - \frac{1}{2}\hat{E} > 0$, $\|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\hat{E}\|}{2}$. Let $(x^k)_{k \in \mathbb{N}}$ be the sequence generated by Distributed PADMM⁺ for any initial point (x^0, y^0) , (x^1, y^1) . Then for all $n \in Q$ the sequence $(x_n^k)_{k \in \mathbb{N}}$ converges to a solution of Problem (6.1).

7.3 A Distributed asynchronous primal-dual splitting algorithm with preconditioning

In this section, we use the randomized coordinate descent on the above algorithm, we call this algorithm as preconditioned distributed asynchronous primal-dual splitting algorithm (PDAPDS). This algorithm has the following attractive property:

Firstly, it significantly accelerates the convergence on problems with irregular D . Moreover, it leaves the computational complexity of the iterations basically unchanged. Finally, if we let $(\zeta^k)_{k \in \mathbb{N}}$ be a sequence of i.i.d. random variables valued in 2^Q . The value taken by ζ^k represents the agents that will be activated and perform a prox on their x

variable at moment k . The asynchronous algorithm goes as follows:

Algorithm 8 PDAPDS.

Initialization: $x^0, x^1 \in \mathcal{X}$, $y^0, y^1 \in \mathcal{Y}$.

Do

- Define $\eta^k = y^k + \alpha_k(y^k - y^{k-1})$,
 $\xi^k = x^k + \alpha_k(x^k - x^{k-1})$,
 - Select a random set of agents $\zeta^{k+1} = \mathcal{B}$.
 - For any $n \in \mathcal{B}$, Agent n performs the following operations :
 - For all $m \sim n$, do

$$y_{\{n,m\}}^{k+1}(n) = \frac{\eta_{\{n,m\}}^k(n) - \eta_{\{n,m\}}^k(m)}{2} + \frac{\xi_n^k - \xi_m^k}{2},$$
 - $x_n^{k+1} = \text{prox}_{\tilde{T}g_n/d_n}[(I - \tilde{T}\Psi^{-1})\xi_n^k - \frac{\tilde{T}}{d_n}\nabla f_n(\xi_n^k) + \frac{\tilde{T}}{d_n}\sum_{m:\{n\sim m\}\in E}(\Psi^{-1}\xi_m^k + \eta_{\{n,m\}}^k(m))]$.
 - For all $m \sim n$, send $\{x_n^{k+1}, y_{\{n,m\}}^{k+1}(n)\}$ to Neighbor m .
 - For any agent $n \notin \mathcal{B}$, $x_n^{k+1} = \xi_n^k$, and $y_{\{n,m\}}^{k+1}(n) = \eta_{\{n,m\}}^k(n)$ for all $m \sim n$.
 - Increment k .
-

Assumption 7.2. The collections of sets $\{\mathcal{B}_1, \mathcal{B}_2, \dots\}$ such that $\mathbb{P}[\zeta^1 = \mathcal{B}_i]$ is positive satisfies $\bigcup \mathcal{B}_i = \mathcal{Q}$.

Theorem 7.2. Assume that the minimization Problem (6.1) is consistent, T and Ψ are two diagonal matrices which have same dimensional. Let Assumption 6.1, Assumption 7.1 and 7.2 hold true, and $\tilde{T}^{-1} - \frac{1}{2}\hat{E} > 0$, $\|\tilde{T}^{-1}\| - \|\Psi^{-1}\| > \frac{\|\hat{E}\|}{2}$. Let $(x_n^k)_{n \in \mathcal{Q}}$ be the sequence generated by PDAPDS for any initial point (x^0, y^0) , (x^1, y^1) . Then the sequence $x_1^k, \dots, x_{|\mathcal{Q}|}^k$ converges to a solution of Problem (6.1).

Proof. Let $(\bar{f}, \bar{g}, h) = (f \circ D^{-1}, g \circ D^{-1}, h)$ where f, g, h and D are the ones defined in the Problem 7.2. By Equations (3.10). We write these equations more compactly as $(y^{k+1}, x^{k+1}) = T(\xi^k, \eta^k)$ where $(\xi^k, \eta^k) = (x^k, y^k) + \alpha_k[(x^k, y^k) - (x^{k-1}, y^{k-1})]$, the operator T acts in the space $\mathcal{Z} = \mathcal{Y} \times \mathcal{R}$, and \mathcal{R} is the image of $\mathcal{X}^{|\mathcal{Q}|}$ by D . Then from the proof of Lemma 3.1, we know that T is \tilde{a} -averaged, where $\tilde{a} = (2 - a_1)^{-1}$ and $a_1 = \frac{\|\hat{E}\|}{2}(\|\tilde{T}^{-1}\| - \|\Psi^{-1}\|)^{-1}$. Defining the selection operator \mathcal{S}_n on \mathcal{Z} as $\mathcal{S}_n(\eta, D\xi) =$

$(\eta_\epsilon(n)_{\epsilon \in Q: n \in \epsilon}, \xi_n)$. So, we obtain that $\mathcal{Z} = \mathcal{S}_1(\mathcal{Z}) \times \cdots \times \mathcal{S}_{|Q|}(\mathcal{Z})$ up to an element reordering. Identifying the set \mathcal{J} introduced in the notations of Section 5.1 with Q , the operator $T^{(\zeta^k)}$ is defined as follows:

$$\mathcal{S}_n(T^{(\zeta^k)}(\eta, D\xi)) = \begin{cases} \mathcal{S}_n(T(\eta, D\xi)), & \text{if } n \in \zeta^k, \\ \mathcal{S}_n(\eta, D\xi), & \text{if } n \notin \zeta^k. \end{cases}$$

Then by Theorem 5.1, we know the sequence $(y^{k+1}, Dx^{k+1}) = T^{(\zeta^{k+1})}(\eta^k, D\xi^k)$ converges almost surely to a solution of Problem (3.11). Moreover, from Lemma 7.1, we have the sequence x^k converges almost surely to a solution of Problem (6.1).

Therefore we need to show that the operator $T^{(\zeta^{k+1})}$ is translated into the PDAPDS algorithm. The definition (7.1) of h shows that

$$h^*(\varphi) = \Sigma_{\epsilon \in E} \iota_{\mathcal{C}_2^\perp}(\varphi_\epsilon),$$

where $\mathcal{C}_2^\perp = \{(x, -x) : x \in \mathcal{X}\}$. Therefore, writing

$$(\zeta^{k+1}, v^{k+1} = Dq^{k+1}) = T(\eta^k, \lambda^k = D\xi^k),$$

then by Eq. (3.10a),

$$\zeta_\epsilon^{k+1} = \text{proj}_{\mathcal{C}_2^\perp}(\eta_\epsilon^k + \Psi^{-1}\lambda_\epsilon^k).$$

Observe that contrary to the case of the synchronous algorithm (7.3), there is no reason here for which $\text{proj}_{\mathcal{C}_2^\perp}(\eta_\epsilon^k) = 0$. Getting back to $(y^{k+1}, Dx^{k+1}) = T^{(\zeta^{k+1})}(\eta^k, \lambda^k = D\xi^k)$, we have for all $n \in \zeta^{k+1}$ and all $m \sim n$,

$$\begin{aligned} y_{\{n,m\}}^{k+1}(n) &= \frac{\eta_{\{n,m\}}^k(n) - \eta_{\{n,m\}}^k(m)}{2} + \frac{\lambda_{\{n,m\}}^k(n) - \lambda_{\{n,m\}}^k(m)}{2} \\ &= \frac{\eta_{\{n,m\}}^k(n) - \eta_{\{n,m\}}^k(m)}{2} + \frac{\xi_n^k - \xi_m^k}{2}. \end{aligned}$$

By Equation (3.10b) we also get

$$v^{k+1} = \arg \min_{w \in \mathcal{R}} [\bar{g}(w) + \langle \nabla \bar{f}(\lambda^k), w \rangle + \frac{\|w - \lambda^k + \tilde{T}(2y^{k+1} - \eta^k)\|_{\tilde{T}^{-1}}^2}{2}].$$

Upon noting that $\bar{g}(D\xi) = g(\xi)$ and $\langle \nabla \bar{f}(\lambda^k), D\xi \rangle = \langle (D^{-1})^* \nabla f(D^{-1}D\xi^k), D\xi \rangle = \langle \nabla f(\xi^k), \xi \rangle$, the above equation becomes

$$q^{k+1} = \arg \min_{w \in \mathcal{X}} [g(w) + \langle \nabla f(\xi^k), w \rangle + \frac{\|D(w - \xi^k) + \tilde{T}(2y^{k+1} - \eta^k)\|_{\tilde{T}^{-1}}^2}{2}].$$

Recall that $(D^*Dx)_n = d_n x_n$. Hence, for all $n \in \zeta^{k+1}$, we get after some computations

$$x_n^{k+1} = \text{prox}_{\tilde{T}g_n/d_n}[\xi_n^k - \frac{\tilde{T}}{d_n} \nabla f_n(\xi_n^k) + \frac{\tilde{T}}{d_n} (D^*(2y^{k+1} - \eta^k))_n].$$

Using the identity $(D^*y)_n = \sum_{m:\{n,m\} \in E} y_{\{n,m\}}(n)$, it can easy check these equations coincides with the x -update in the PDAPDS algorithm. □

8 Numerical experiments

We consider the problem of l_1 -regularized logistic regression. Denoting by m the number of observations and by q the number of features, the optimization problem writes

$$\inf_{x \in \mathbb{R}^q} \frac{1}{m} \sum_{i=1}^m \log(1 + e^{-y_i a_i^T x}) + \tau \|x\|_1, \quad (8.1)$$

where the $(y_i)_{i=1}^m$ are in $\{-1, +1\}$, the $(a_i)_{i=1}^m$ are in \mathbb{R}^q , and $\tau > 0$ is a scalar. Let $(\mathcal{W})_{n=1}^N$ indicate a partition of $\{1, \dots, m\}$. The optimization problem then writes

$$\inf_{x \in \mathbb{R}^q} \sum_{n=1}^N \sum_{i \in \mathcal{W}_n} \frac{1}{m} \log(1 + e^{-y_i a_i^T x}) + \tau \|x\|_1, \quad (8.2)$$

or, splitting the problem between the batches

$$\inf_{x \in \mathbb{R}^{Nq}} \sum_{n=1}^N \left(\sum_{i \in \mathcal{W}_n} \frac{1}{m} \log(1 + e^{-y_i a_i^T x_n}) + \frac{\tau}{N} \|x_n\|_1 \right) + \iota_{\mathcal{C}}(x), \quad (8.3)$$

where $x = (x_1, \dots, x_N)$ is in \mathbb{R}^{Nq} . It is easy to see that Problems (8.1), (8.2) and (8.3) are equivalent and Problem (8.3) is in the form of (6.2).

9 Conclusion

In this paper, we introduced a new framework for stochastic coordinate descent and used on a algorithm called ADMMSD⁺. As a byproduct, we obtained a stochastic

approximation algorithm with dynamic stepsize which can be used to handle distinct data blocks sequentially. We also obtained an asynchronous distributed algorithm with dynamic stepsize which enables the processing of distinct blocks on different machines.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (11131006, 41390450, 91330204, 11401293), the National Basic Research Program of China (2013CB 329404), the Natural Science Foundations of Jiangxi Province (CA201107114, 20114BAB 201004).

References

- [1] L. Condat, A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms, *Journal of Optimization Theory and Applications*, vol. 158, no. 2, pp. 460C479, 2013.
- [2] Bianchi P, Hachem W and Iutzeler F 2014 A Stochastic coordinate descent primal-dual algorithm and applications to large-scale composite (arXiv:1407.0898v1 [math.OC] 3 Jul 2014) *Optimization*
- [3] Bauschke, H.H., Combettes, P.L.: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York (2011)
- [4] Ogura, N., Yamada, I.: Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping. *Numer. Funct. Anal. Optim.* 23(1C2), 113-137 (2002)
- [5] Nesterov, Yu.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR* 269(3), 543-547 (1983)
- [6] Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* 2(1), 183-202 (2009)
- [7] Nesterov, Y.: *Introductory lectures on convex optimization: a basic course*. In: *Applied Optimization*, vol. 87. Kluwer Academic Publishers, Boston, MA (2004)

- [8] Tseng, P.: On accelerated proximal gradient methods for convexconcave optimization. Technical report (2008)
- [9] BoT, R. I., Csetnek, E. R. and Hendrich, C.:Inertial Douglas-Rachford splitting for monotone inclusion problems. arXiv:1403.3330v2 [math.OC] 28 Mar 201
- [10] I. Daubechies, M. Defrise, and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on pure and applied mathematics, vol. 57, no. 11, pp. 1413-1457, 2004.
- [11] M. Bacak, The proximal point algorithm in metric spaces, Israel Journal of Mathematics, vol. 194, no. 2, pp. 689-701, 2013.
- [12] Pock, T., Chambolle, A.: Diagonal preconditioning for first order primal-dual algorithms. In: Proceedings of the International Conference of Computer Vision (ICCV 2011), pp. 1762-1769 (2011)
- [13] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological), pp. 267C288, 1996.
- [14] F. Alvarez, On the minimizing property of a second order dissipative system in Hilbert spaces, SIAM Journal on Control and Optimization 38(4), 1102-1119, 2000
- [15] F. Alvarez, Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space, SIAM Journal on Optimization 14(3), 773-782, 2004
- [16] F. Alvarez, H. Attouch, An inertial proximal method for maximal monotone operators via discretization of a nonlinear oscillator with damping, Set-Valued Analysis 9, 3-11, 2001
- [17] Forero, P A, Cano A and Giannakis G B 2010 Consensus-based distributed support vector machines The Journal of Machine Learning Research 99 1663-1707.
- [18] Agarwal A, Chapelle O, Dudík M, and Langford J 2011 A reliable effective terascale linear learning system arXiv preprint arXiv:1110.4198.

- [19] P. Bianchi and J. Jakubowicz, Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization, *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 391- 405, February 2013.
- [20] S.S. Ram, V.V. Veeravalli, and A. Nedic, Distributed and recursive parameter estimation in parametrized linear state-space models, *IEEE Trans. on Automatic Control*, vol. 55, no. 2, pp. 488-492, 2010.
- [21] P. Bianchi, G. Fort, and W. Hachem, Performance of a distributed stochastic approximation algorithm, *IEEE Transactions on Information Theory*, 59(11) (2013), 7405 C 7418 .
- [22] Moudafi, A., Oliny, M.: Convergence of a splitting inertial proximal method for monotone operators. *J. Comput. Appl. Math.* 155, 447- 454 (2003)
- [23] Yu. Nesterov, Efficiency of coordinate descent methods on huge-scale optimization problems, *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341-362, 2012.
- [24] O. Fercoq and P. Richtarik, Accelerated, parallel and proximal coordinate descent, *arXiv preprint arXiv:1312.5799*, 2013.